We construct efficient estimators of the identifiable parameters in a regression model when the errors follow a stationary parametric ARCH($P$) process. We do not assume a functional form for the conditional density of the errors, but do require that it be symmetric about zero. The estimators of the mean parameters are adaptive in the sense of Bickel [2]. The ARCH parameters are not jointly identifiable with the error density. We consider a reparameterization of the variance process and show that the identifiable parameters of this process are adaptively estimable.

1. INTRODUCTION

We consider the problem of obtaining efficient estimators of the identifiable parameters in the following linear regression model where the errors are conditionally heteroskedastic according to an ARCH($P$) process:

$$y_t = \beta^T x_t + u_t; \quad u_t = \epsilon_t \sigma_t; \quad (1.1)$$

$$\sigma_t^2 = a + \sum_{j=1}^{P} c_j u_{t-j}^2, \quad t = 1,2,\ldots,T. \quad (1.2)$$

This specification of the error process was originally suggested in Engle [10], and was employed there to model United Kingdom inflation rates. It has been used in countless empirical studies—see the survey papers of Engle and Bollerslev [12] and Bollerslev, Chou, and Kroner [7] for references.

The ARCH specification rationalizes two well-established empirical regularities about financial and macroeconomic time series. When $c_j$, $j = 1,2,\ldots,P$ are all positive, the process $\sigma_t^2$ is positively serially dependent. This is an important feature: Many financial and macroeconomic time series are characterized by episodic bursts of volatility followed by more tranquil periods. Uncertainty about future events—and the consequent risk to investors—varies over time yet typically is closely related to previous assessments of uncertainty.

In addition, even when $\epsilon_t$ are i.i.d. normal, the unconditional distribution of the innovation $u_t$ will be leptokurtic by virtue of the mixing that random

I would like to thank Paul Ruud, Tom Rothenberg, Doug Steigerwald, and Peter Bickel for helpful discussions. The comments of a co-editor and two referees helped improve the presentation. Any remaining errors are my own.
\( \sigma \) induces. This is consistent with the findings of many researchers—for example, Mandelbrot [28] and Gallant et al. [17]—who have found that the distribution of stock returns tends to have heavier tails than the normal.

A number of generalizations of the ARCH specification are given in Engle and Bollerslev [12] and Bollerslev, Chou, and Kroner [7]. Recent research has focused on Generalized ARCH and Integrated GARCH models—see Bollerslev [5] and Lumsdaine [27], on exponential ARCH/GARCH models; see Geweke [18] and Nelson [30]; and on semiparametric ARCH/GARCH modelling, see Engle and Gonzalez-Rivera [13] and Whistler [45]. A number of authors have proposed alternatives to the ARCH paradigm. In particular, Spanos [39] and Shephard [38] highlight deficiencies in the ARCH modelling approach and suggest alternatives.

In this paper we are concerned with the semiparametric approach. Specifying a parametric model for the density of \( \epsilon \), imposes strong restrictions on the data generation process, especially when the normal distribution is used. Although Bollerslev [5] and Weiss [44] show that the Gaussian (Pseudo) Maximum Likelihood Estimator (hereafter the PMLE) of \( \theta = (\beta^T, a, c^T)^T \), where \( c = (c_1, c_2, \ldots, c_p)^T \), is \( \sqrt{T} \) consistent asymptotically normal under quite general conditions on the error density \( f \), this estimator is inefficient for non-normal \( f \)'s. With the large sample sizes available for financial data one ought to be able to do better.

Engle and Gonzalez-Rivera [13] consider a semiparametric extension of (1.1) and (1.2). They retained the linear relationship (1.2) yet allowed \( f(\cdot) \) to be of unknown form. They used nonparametric estimates of the score function of \( \epsilon \) to estimate the parameters of a GARCH process. They report Monte-Carlo simulation results that suggest improvements for this method over the Gaussian PMLE when the true error distribution was non-normal.

We examine further this semiparametric model. The issues we address here are twofold. First, what is the information bound for estimation of \( \theta \) when no parametric structure is assumed for the error density? In particular, is this an adaptive situation—can one in principle estimate \( \theta \) as well when \( f \) is unknown as when it is known? Second, is it possible to construct an estimate of \( \theta \) that achieves the information bound asymptotically?

Bickel [2] gives a necessary condition for adaptation in the context of a semiparametric model \( P_{\theta,G} \), where \( \theta \) is a finite dimensional parameter and \( G \) is an infinite dimensional nuisance parameter: The scores for \( \theta \) must be orthogonal to the tangent space for \( G \). In particular, the scores for \( \theta \) must be orthogonal to the scores for the scalar parameter \( \tau \) for each parameterization \( G(\cdot; \tau) \) of \( G(\cdot) \).

This orthogonality condition is satisfied in a number of semiparametric models. In particular, Bickel [2] shows that the information bound for estimating the slope parameters in a linear regression is the same whether or not the error density is known. Kreiss [21,22] extends these results to a time se-
ries context. He shows that it is possible to estimate the identifiable parameters of a stationary invertible ARMA model adaptively in the presence of an unknown error density. His results are considerably easier to derive when the error distribution is symmetric.

Engle and Gonzalez-Rivera [13] examined the performance of their semiparametric estimator when the true error density was either $t_5$ or gamma distributed. They found that although the estimator generally outperformed the Gaussian PMLE, it appeared to be considerably less efficient than the MLE. They suggested that the semiparametric estimator was not adaptive even when the error density was symmetric.

In this paper, we consider only the situation where the unknown error density is symmetric about zero. We find that under the specification (1.1) and (1.2) the mean parameters $\beta$ can be estimated adaptively when the error density is unknown, while the parameters $(a, c, f)$ are not jointly identifiable. To deal with the identifiability problem we reparameterize (1.2) as

$$\sigma^2(\theta) = e^\alpha \left[ 1 + \sum_{j=1}^P \gamma_j u_{t-j}^2 \right].$$

(1.3)

This parameterization separates the overall scale effect ($e^\alpha$) from the relative effects measured by $\gamma_j$, $j = 1, 2, \ldots, P$. We find that the parameter $\alpha$ has zero information when $f$ is unknown, while the scores for $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_P)^T$ are orthogonal to the tangent space for the unknown error density, that is, $\gamma$ is in principle adaptively estimable.

We construct an estimator of the identifiable parameters and show that our estimator is asymptotically equivalent to the MLE, and hence is adaptive. To establish the asymptotic properties of our estimator we use the methodology developed in Le Cam [25], Bickel [2], Kreiss [21,22], and Swensen [43].

Engle and Gonzalez-Rivera [13] employ a different approach to ensuring semiparametric identifiability—they normalize the variance of $f$ to be one. This introduces a nonlinear constraint on the class of allowable error densities which is difficult to incorporate in information bound calculations. It is not clear whether this approach will give rise to adaptively estimable parameters other than $\beta$, although the preliminary calculations presented in Steigerwald [41] suggest it may not.

The paper is structured as follows. In Section 2 we examine whether Bickel’s orthogonality condition holds in the original ARCH model and in our reparameterization. In Section 3 we state our assumptions. In Section 4 we establish the fundamental LAN property for our reparameterized ARCH model. In Section 5 we establish properties of linearized MLE’s of the unknown parameters when the error density is known. In Section 6 we construct an estimator that does not require knowledge of the error density and is adaptive. Section 7 concludes.
2. IS ADAPTIVE ESTIMATION POSSIBLE?

2.1. The Location Scale Model

We first review the theory developed in Bickel and Ritov [3] concerning information bounds in semiparametric models in the context of the location scale model

\[ y = \mu + \varepsilon \sigma, \]

where \( \varepsilon \) is distributed symmetrically about zero with density \( f \). When \( f \) is known, the scores for the unknown parameters \((\mu, \sigma)\) are

\[
\hat{\ell}_\mu(\varepsilon) = -\sigma^{-1} \frac{f'}{f}(\varepsilon) = -\sigma^{-1} \psi_1(\varepsilon);
\]

\[
\hat{\ell}_\sigma(\varepsilon) = -\sigma^{-1} \left[ \varepsilon \frac{f'}{f}(\varepsilon) + 1 \right] = -\sigma^{-1} \psi_2(\varepsilon).
\]

These scores are mutually orthogonal when \( f \) is symmetric about zero. In this case the information bounds for estimating \( \mu \) and \( \sigma \) when \( f \) is known are given by \( \sigma^2 I_1(f)^{-1} \) and \( \sigma^2 I_2(f)^{-1} \) respectively, where

\[
I_1(f) = E[\psi_1(\varepsilon)^2], \quad I_2(f) = E[\psi_2(\varepsilon)^2].
\]

We verify Bickel's orthogonality condition using the following heuristic argument. Suppose that \( f \) is parameterized by a scalar parameter \( \tau \) such that \( f(\cdot; \tau) \) is symmetric about zero for all \( \tau \), and let \( f_\tau(\cdot; \tau) \) denote the partial derivative of \( f(\cdot; \tau) \) with respect to \( \tau \). Since \( f(\cdot; \tau) \) is symmetric about zero for all \( \tau \),

\[
f_\tau(\cdot; \tau) = \lim_{\delta \to 0} \frac{f(\cdot; \tau + \delta) - f(\cdot; \tau)}{\delta}
\]

is also symmetric about zero. The score function for \( \tau \) in the parametric model \( P_\theta \), where \( \theta = (\mu, \sigma, \tau)^T \), is \( \hat{\ell}_\tau \), where

\[
\hat{\ell}_\tau(\varepsilon) = \frac{f_\tau}{f}(\varepsilon),
\]

and is therefore symmetric about zero. Furthermore, \( \hat{\ell}_\tau \) is orthogonal to \( \hat{\ell}_\mu \), and the information bound for estimating \( \mu \) in the presence of the unknown parameter \( \tau \) is \( I_1(f) \) — knowledge of \( \tau \) provides no useful information about \( \mu \). Since the parameterization was arbitrary, we conclude that knowledge of the error density \( f \) is irrelevant as far as estimation of \( \mu \) is concerned. However, the parameter \( \sigma \) is not identifiable without further restrictions on \( f \).

Suppose that instead of the natural parameterization \((\mu, \sigma)\), one had parameterized the model by \((\xi, \eta)\), where \( \xi = (\mu + \sigma)/2 \) and \( \eta = (\mu - \sigma)/2 \). In this case, the scores for \( \xi \) and for \( \eta \) are both correlated with those for \( \tau \),
and one might conclude that this was not an adaptive situation.¹ We suggest that the ARCH model manifests this phenomenon—in (1.2) each of the parameters \( a \) and \( c_j, j = 1, 2, \ldots, p \) gives information about scale.

2.2. Engle’s ARCH Model

We now examine the ARCH model defined by (1.1) and (1.2). Let \( x_t \) be a \( K \) by 1 vector of fixed regressors, and suppose that \( \varepsilon_t \) is i.i.d. zero mean, with density \( f \) symmetric about zero. Let \( (\beta^T, a, c^T)^T = \phi \), where \( c = (c_1, c_2, \ldots, c_p)^T \) and let the zero subscript denote the true parameter value where necessary. Furthermore, suppose that the initial conditions \( Y_0 = (\varepsilon_0, \ldots, \varepsilon_{1-p}, \sigma_0^2, \ldots, \sigma_{1-p}^2) \) are observed, and let \( f_0(Y_0; \phi) \) denote the unconditional density of \( Y_0 \).

The sample log likelihood \( \ell(Y_0; Y_1, \ldots, Y_T; \phi) \) for the ARCH model (1.1) and (1.2) can be written as

\[
\ell = \log f_0(Y_0; \phi) + \sum_{t=1}^{T} \log f(\varepsilon_t(\phi)) - \frac{1}{2} \sum_{t=1}^{T} \log [\sigma_t^2(\phi)],
\]

where

\[
e_t(\phi) = \frac{(y_t - \beta^T x_t)}{\sigma_t(\phi)}; \quad \sigma_t^2(\phi) = a + \sum_{j=1}^{p} c_j (y_{t-j} - \beta^T x_{t-j})^2.
\]

We shall assume that the process \( \sigma_t^2(\phi_0) \) is stationary, and that \( f_0 \) makes a vanishingly small contribution to the asymptotic properties of the MLE. We focus our attention on the conditional likelihood that drops \( f_0 \). Let \( \delta_t(\phi) = (\delta_{t\beta}, \delta_{ta}, \delta_{tc})^T \) denote the \( K + P + 1 \) vector of period \( t \) contributions to the sample scores of the conditional likelihood, and let \( \hat{\ell}_{T\phi} = (\hat{\ell}_{T\beta}, \hat{\ell}_{Ta}, \hat{\ell}_{Tc})^T \), where \( \hat{\ell}_{T\phi}(\phi) = \sum_{t=1}^{T} \delta_t(\phi) \). Then

\[
\hat{\ell}_{T\beta}(\phi) = -\sum_{t=1}^{T} \left\{ \sigma_t(\phi)^{-1} x_t \psi_1(\varepsilon_t(\phi)) + W_t(\phi) \psi_2(\varepsilon_t(\phi)) \right\} = \hat{\ell}_{T\beta 1}(\phi) + \hat{\ell}_{T\beta 2}(\phi),
\]

where

\[
W_t(\phi) = -\sum_{j=1}^{P} c_j x_{t-j} \frac{(y_{t-j} - \beta^T x_{t-j})}{\sigma_t^2(\phi)} = -\sum_{j=1}^{P} c_j x_{t-j} w_{t-j}(\phi),
\]

and

\[
w_{t-j}(\phi) = \frac{(y_{t-j} - \beta^T x_{t-j})}{\sigma_t^2(\phi)}.
\]

Both \( W_t(\phi) \) and \( \sigma_t^2(\phi) \) depend only on the past. Similarly,

\[
\hat{\ell}_{Tc}(\phi) = -\frac{1}{2} \sum_{t=1}^{T} \psi_2(\varepsilon_t(\phi)) \psi_3(\phi); \quad \hat{\ell}_{Ta}(\phi) = -\frac{1}{2} \sum_{t=1}^{T} \psi_2(\varepsilon_t(\phi)) \sigma_t(\phi)^{-2},
\]
where \( v_t = (v_{1t}, v_{2t}, \ldots, v_{pt})^T \), and
\[
v_{ji}(\phi) = (y_{t-j} - \beta^T x_{t-j})^2/\sigma_j^2(\phi), \quad j = 1, 2, \ldots, P
\]
which also depends only on the past.

To investigate whether Bickel’s orthogonality conditions are met in the ARCH model, we proceed heuristically as in Section 2.1. Suppose that \( f \) is parameterized by a scalar parameter \( \tau \). In this case,
\[
\hat{\theta}_T = \sum_{t=1}^T \frac{f_e}{f}(\varepsilon_t),
\]
where \((f_e/f)(\cdot)\) is symmetric about zero. Although \( \delta_{\beta z} \) is an even function of \( \varepsilon_t \), it is orthogonal to \( \delta_{\varepsilon} \). Since \( W_t \) is independent of \( \varepsilon_t \) and is mean zero, \( \hat{\theta}_T \) and \( \hat{\theta}_T \) are mutually orthogonal. Thus there is no efficiency loss from not knowing \( \tau \). Since we only exploit symmetry in obtaining this orthogonality, this result carries over to the semiparametric model. One should be able to construct adaptive estimates of \( \beta \), provided one can estimate the score functions \( \psi_1 \) suitably well.

This orthogonality does not hold for the remaining parameters, since \( \hat{\theta}_T \), \( \hat{\theta}_T \), and \( \hat{\theta}_T \), are in general correlated. We argue that this correlation is a manifestation of the fact that we cannot separately identify \( (a, c, f) \). Before discussing information bounds in this model, we must deal with this issue.

We reparameterize the ARCH process according to (1.3), that is,
\[
\sigma^2(\theta) = e^\alpha \left[ 1 + \sum_{j=1}^P \gamma_j (y_{t-j} - \beta^T x_{t-j})^2 \right],
\]
where \( \theta = (\beta^T, \alpha, \gamma^T)^T \), and \( \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_P)^T \). Now let \( \hat{\theta}_T(\theta) = \sum_{t=1}^T \delta_{\theta}(\theta) \), where \( \delta_{\theta}(\theta) = (\delta^T_{\beta z}, \delta_{\varepsilon \alpha}, \delta_{\varepsilon \gamma})^T \) and \( \hat{\theta}_T = (\hat{\theta}_T, \hat{\theta}_T, \hat{\theta}_T)^T \). Then
\[
\hat{\theta}_T = -\sum_{t=1}^T \left\{ \sigma_j(\theta)^{-1} x_t \psi_1(\varepsilon_t(\theta)) + \overline{W}_t(\theta) \psi_2(\varepsilon_t(\theta)) \right\}
\]
\[
= \hat{\theta}_T + \hat{\theta}_T(\theta),
\]
where
\[
\overline{W}_t(\theta) = -\sum_{j=1}^P \gamma_j x_{t-j} \frac{(y_{t-j} - \beta^T x_{t-j})}{\sigma_j^2(\theta)} = -\sum_{j=1}^P \gamma_j x_{t-j} \overline{w}_{t-j}(\theta),
\]
\[
\overline{\sigma}_j^2(\theta) = 1 + \sum_{j=1}^P \gamma_j (y_{t-j} - \beta^T x_{t-j})^2; \quad \overline{w}_{t-j}(\theta) = \frac{(y_{t-j} - \beta^T x_{t-j})}{\sigma_j^2(\theta)};
\]
while
\[
\hat{\theta}_T(\theta) = -\frac{1}{2} \sum_{t=1}^T \psi_2(\varepsilon_t(\theta)); \quad \hat{\theta}_T(\theta) = -\frac{1}{2} \sum_{t=1}^T \psi_2(\varepsilon_t(\theta)) \hat{\psi}_1(\theta),
\]
where

\[ \hat{\nu}_j(\theta) = \frac{(y_{t-j} - \beta^T x_{t-j})^2}{\hat{\sigma}^2(\theta)}, \quad j = 1, 2, \ldots, P. \]

Notice that \(\omega_i(\theta_0)^{-1}, \tilde{\nu}_i(\theta_0),\) and \(\hat{\nu}_i(\theta_0)\) are all stationary ergodic processes, and are all bounded from above when \(\alpha_0 > -\infty\) and \(\gamma_j \geq 0, j = 1, 2, \ldots, P.\)

The efficient score function for \(\gamma\) in the presence of unknown \(\alpha\), obtained by projecting \(\hat{\nu}_\gamma\) onto \(\hat{\nu}_\alpha\), is

\[ \hat{\nu}_\gamma = \hat{\nu}_\gamma - E[\hat{\nu}_\gamma \hat{\nu}_\alpha] E[\hat{\nu}_\alpha^{-1} \hat{\nu}_\alpha - \frac{1}{2} \sum_{i=1}^{T} \psi_2(\epsilon_i)(\tilde{\nu}_i - \tilde{\nu}), \]

where \(\tilde{\nu} = E[\tilde{\nu}_i].\) Now \(\hat{\nu}_\gamma^*\) is orthogonal to any score function \(\hat{\nu}_\gamma\), where

\[ \hat{\nu}_\gamma = \sum_{i=1}^{T} \frac{f(\epsilon_i)}{f}\epsilon_i, \]

for \(f(\cdot; \tau)\) any parameterization of the symmetric function \(f(\cdot).\) Therefore, the information bound for estimating \(\gamma\) is the same whether or not \(f\) is known. Under suitable regularity conditions we should be able to estimate \(\gamma\) adaptively.

In the sequel we construct an estimator of \((\beta^T, \gamma^T)^T\) that achieves the information bound provided that \(f\) is symmetric about zero.

Remark. Consider the exponential ARCH model

\[ \log[\sigma_j^2] = \alpha + \sum_{j=1}^{P} \gamma_j r(\epsilon_{t-j} \sigma_{t-j}), \quad \text{(2.3)} \]

where \(r(\cdot)\) is a known function, see Nelson [30]. If \(E[r'(\epsilon, \sigma)] = 0,\) the scores for \(\beta\) are orthogonal to those for \(f.\) In this case, we argue in Appendix A that both \(\beta\) and \(\gamma\) are in principle adaptively estimable, although see Bickel and Ritov [3] for a cautionary tale in this regard.

3. ASSUMPTIONS

Although (1.2) and (1.3) generate the same family of probability measures, the relevant parameter spaces differ. To avoid any ambiguity we shall restrict our attention to parameterization (1.3). We use the following conditions.

Assumption 1. The random variables \(\{\epsilon_t\}\) are i.i.d., with absolutely continuous Lebesgue density \(f,\) where \(f(x) > 0 \forall x \in \mathbb{R}.\)

Let the score functions \(\psi_1\) and \(\psi_2\) be defined as follows:

\[ \psi_1(x) = \frac{f'}{f}(x); \quad \psi_2(x) = x \frac{f'}{f}(x) + 1, \]
where $\frac{1}{2}f^{-1/2}(x)f'(x)$ is the quadratic mean derivative of $f(x)^{1/2}$, that is,
\[ \lim_{m \to 0} \frac{1}{m^2} \int \left[ f(x + m)^{1/2} - f(x)^{1/2} - \frac{m}{2} f^{-1/2}(x)f'(x) \right]^2 \, dx = 0. \quad (3.1) \]

We do not assume that $f$ is necessarily differentiable everywhere in the usual sense, although the following assumptions restrict the lack of smoothness that can be permitted.

Assumption 2. The density $f$ has finite Fisher information for both scale and location parameters,
\[ 0 < I_1(f) = \int \psi_1(x)^2 f(x) \, dx < \infty; \quad 0 < I_2(f) = \int \psi_2(x)^2 f(x) \, dx < \infty. \]

Assumption 3. The score functions, $\psi_i$, $i = 1, 2$, satisfy the following conditions:
\begin{enumerate}
  \item $\int [\psi_i((x + m)/(1 + s)) - \psi_i(x)]^2 f(x) \, dx = 0$ as $m, s \to 0$, $i = 1, 2$,
  \item $\int m^{-1} \psi_1((x + m)/(1 + s)) f(x) \, dx = -I_1(f)$ as $m, s \to 0$,
  \item $\int s^{-1} \psi_2((x + m)/(1 + s)) f(x) \, dx = -I_2(f)$ as $m, s \to 0$,
  \item $\int s^{-1} \psi_1((x + m)/(1 + s)) f(x) \, dx = 0$ as $m, s \to 0$,
  \item $\int m^{-1} \psi_2((x + m)/(1 + s)) f(x) \, dx = 0$ as $m, s \to 0$.
\end{enumerate}

Remark. These are essentially second derivative conditions, and are satisfied by a large class of densities: for example, the normal, the GED distribution considered in Nelson [30], and the Laplace distribution. Assumption 3(1) is an obvious extension of Assumption 5(1) in Kreiss [21], while Assumption 3(2) is A5(ii) of Kreiss [21].

Remark. Lind and Roussas [26] establish in a more general context that quadratic mean differentiability assumptions such as (3.1) and Assumption 3 imply Cramer’s conditions (see Cramer [9], p. 500).

Assumption 4. The error density also satisfies:
\begin{enumerate}
  \item The density $f$ is symmetric about zero,
  \item $\int x^2 f(x) \, dx < \infty$,
  \item $\int \psi_j(x)^4 f(x) \, dx < \infty$, \quad $j = 1, 2$.
\end{enumerate}

Assumption 5. The parameter space $\Theta$ is an open subset of $\mathbb{R}^{K+P+1}$ that satisfies various restrictions such that
\begin{enumerate}
  \item The process $\{\sigma_t^2\}_{t=1}^\infty$ is bounded below by a constant $\sigma > 0$.
  \item The process $\{\sigma_t^2\}_{t=1}^\infty$ is strictly stationary and ergodic.
  \item The process $\{\sigma_t^2\}_{t=1}^\infty$ satisfies $E[\sigma_t^4] < \infty$.
\end{enumerate}

Remark. A sufficient condition for Assumption 5(1) to hold is that $\gamma_1 \geq 0, \gamma_2 \geq 0, \ldots, \gamma_P \geq 0$. 

Nelson and Cao [32] show that these conditions can be weakened somewhat. Primitive conditions on \( \alpha \) and \( \gamma \) and on the distribution of the white noise error that imply Assumption 5(2) are given in Nemec and Linnell [33]. Similar conditions are given in Nelson [31], Sampson [37], and Bougerol and Picard [8] for the GARCH(1,1) model. Assumption 5(3) also requires substantial restrictions on the parameter space as discussed in Bollerslev [5] and Milhoj [29].

For any \( \theta \in \Theta \), let \( P_{T, \theta} \) be the joint probability measure of a sample \( \{ y_t, x_t \}_{t=1}^T \). In the sequel, unless otherwise stated, we let \( P_{T, \theta} \) denote convergence in probability under \( P_{T, \theta_0} \), while \( O_P(\cdot) \) and \( O_\propto(\cdot) \) will also hold under \( P_{T, \theta_0} \). Likewise, \( \Rightarrow \) denotes weak convergence of the associated probability measure under \( P_{T, \theta_0} \). We make an additional assumption:

Assumption 6. The density \( f_\theta(Y_0; \theta) \) is continuous in probability: Let \( \theta_T = \theta_0 + T^{-1/2} h \), and assume that for any \( h \in \mathbb{R}^{K+P+1} \) and \( \forall \theta_0 \in \Theta \),

\[
f_\theta(Y_0; \theta_T) \overset{P}{\Rightarrow} f_\theta(Y_0; \theta_0) \quad \text{as} \quad T \to \infty.
\]

We assume throughout that the \( K \) by 1 vector of explanatory variables \( x_t \) are strictly exogenous, and we therefore condition our inference on \( \{ x_t \}_{t=1}^T \). We define the sequence of \( K \) by \( K \) matrices:

\[
M_T(s) = \{ m_T(s)_{jk} \} = T^{-1} \sum_{t=s+1}^T x_t x_t^T, \quad s = 0, 1, 2, \ldots, P.
\]

We make the following assumption about the regressors:

Assumption 7. The matrix \( M_T(0) \) converges to a finite limit \( M(0) \), where \( M(0) \) is strictly positive definite.

Remark. This assumption on the regressors could be relaxed to allow trending regressors, for example, by assuming Grenander's conditions. In this case, we must replace the \( \sqrt{T} \) norming of our estimator by a suitable matrix as in Swensen [43].

Finally, we shall assume that there exists a \( \sqrt{T} \) consistent estimator \( \hat{\theta}_T \) of \( \theta \). Recall that \( a = e^\alpha \) and \( c_j = e^\alpha \gamma_j \), \( j = 1, 2, \ldots, P \). Weiss [44] and Lumsdaine [27] give conditions under which least-squares estimators and Gaussian PMLE's of the parameters \( a \) and \( c \) are \( \sqrt{T} \) consistent. A delta method argument can then be used to establish the \( \sqrt{T} \) consistency of the resulting estimators of \( \theta \). These authors impose additional moment conditions of various types.

4. LOCAL ASYMPOTIC NORMALITY

In this section, we establish that the log-likelihood ratio of the ARCH model (1.1) and (1.3) satisfies the Local Asymptotic Normality (LAN) condition defined in Theorem 1 below. This condition, introduced in Le Cam [23], con-
trols the behavior of the log-likelihood ratio in a small neighborhood of the true value, requiring that in large samples it be approximately quadratic in a neighborhood of the true parameter. This regularity is essential when establishing the properties of the Newton-Raphson estimators we consider in later sections.

Le Cam [23], Swensen [42,43], and Roussas [36] give conditions under which the log-likelihood ratio of a general stochastic process satisfies the LAN condition. These conditions have been verified for stationary invertible ARMA processes in Kreiss [21], and for linear regression models with autoregressive errors in Swensen [43]. This latter result was extended by Steigerwald [40] to linear regression models with ARMA errors. Generalizations of this concept to Locally Asymptotically Mixed Normal (LAMN) considered in Swensen [42] have found applications in the theory of nonstationary processes—see Phillips [34].

The parameters of interest in the above examples are all location parameters. In the ARCH model, parameters that determine the scale of the process are also of interest. We verify the conditions of Swensen [42] below, using some modifications of the argument presented in Swensen [43].

We first establish some notation. Define the square root of the likelihood ratio \( \lambda \) to be

\[
\lambda(\theta_0, \theta) = \left[ \frac{f_0(Y_0; \theta)}{f_0(Y_0; \theta_0)} \right]^{1/2} \prod_{t=1}^{T} \left[ \frac{f_0(\varepsilon_t(\theta))\sigma_0(\theta_0)}{f_0(\varepsilon_t(\theta_0))\sigma_t(\theta)} \right]^{1/2}.
\]

The log-likelihood ratio is defined as

\[
\Lambda_T(\theta_0, \theta) = \log \left\{ \frac{f_0(Y_0; \theta)}{f_0(Y_0; \theta_0)} \right\} + 2 \sum_{t=1}^{T} \log \phi_t(\theta_0, \theta)
\]

where

\[
\phi_t(\theta_0, \theta) = \left[ \frac{f(\varepsilon_t(\theta))\sigma_t(\theta_0)}{f(\varepsilon_t(\theta_0))\sigma_t(\theta)} \right]^{1/2}.
\]

Let \( S_{T\theta}(\theta) = T^{-1/2} \hat{\varepsilon}_{T\theta} = T^{-1/2} \sum_{t=1}^{T} \delta_{t\theta} \), where \( \hat{\varepsilon}_{T\theta} \) are defined in (2.1) and (2.2). Furthermore, let \( S_{T\beta} = S_{T\beta_1} + S_{T\beta_2}, \) \( S_{T\alpha}, \) and \( S_{T\gamma} \) denote the corresponding subvectors. We now define the information matrix.

**DEFINITION.** Let the information matrix \( J_{T\theta}(\theta_0) \) be the probability limit under \( P_{T, \theta_0} \) of the observed information matrix

\[
J_{T\theta}(\theta_0) = T^{-1} \sum_{t=1}^{T} \delta_{t\theta}(\theta_0)\delta_{t\theta}(\theta_0)^T.
\]

The matrix \( J_{T\theta}(\theta) \) exists by Assumptions 1, 2, 4, 5, and 7; it is strictly positive definite under Assumption 4(b) and 5(3)—see Weiss [44], Lemma 3.2. It has the following structure:
\[
J_{\theta} = \begin{bmatrix}
J_{\beta\beta} & 0 & 0 \\
0 & J_{\alpha\alpha} & J_{\gamma\alpha} \\
0 & J_{\alpha\gamma} & J_{\gamma\gamma}
\end{bmatrix},
\]
(4.1)

where

\[
J_{\beta\beta} = \{\bar{g}_{\beta\beta 1}I_1(f) + \bar{g}_{\beta\beta 2}I_2(f)\}, \quad M(0) = J_{\beta\beta 1} + J_{\beta\beta 2},
\]

\[
J_{\alpha\alpha} = \frac{1}{4} \bar{g}_{\alpha\alpha}I_2(f); \quad J_{\gamma\gamma} = \frac{1}{4} \bar{g}_{\gamma\gamma}I_2(f); \quad J_{\alpha\gamma} = \frac{1}{4} \bar{g}_{\alpha\gamma}I_2(f),
\]

while

\[
g_{t\beta\beta 1}(\theta) = \sigma_t(\theta)^{-2}; \quad g_{t\beta\beta 2}(\theta) = \sum_{j=1}^{P} \gamma_j^2 \tilde{w}_{t-j}(\theta)^2;
\]

\[
g_{t\alpha\alpha}(\theta) = 1; \quad g_{t\gamma\gamma} = \tilde{\nu}_t(\theta)\tilde{\nu}_t(\theta)^T; \quad g_{t\alpha\gamma}(\theta) = \tilde{\nu}_t(\theta),
\]

and \(\bar{g}_{ij} = E_{\theta_0}[g_{ij}(\theta_0)]\) for each \(j\).

With these definitions, we now state the main theorem of this section.

THEOREM 1 (LOCAL ASYMPTOTIC NORMALITY). Assume that Assumptions 1-7 hold, and let \(\theta_T = \theta_0 + T^{-1/2}h\) for any \(h \in \mathbb{R}^{K+p+1}\). Then

1. \(\Lambda_T(\theta_0, \theta_T) - h^T S_T(\theta_0) + \frac{1}{2} h^T J_{\theta\theta}(\theta_0, f) h \xrightarrow{p} 0, \) as \(T \to \infty\).
2. \(S_T(\theta_0) \Rightarrow N(0, J_{\theta\theta}(\theta_0, f))\).

Remark. The asymptotic normality of \(S_T(\theta_0)\) under \(P_{T, \theta_0}\) is easy to establish because \(\{\delta_{\theta\theta}(\theta_0)\}_{t=1}^{\infty}\) is a sequence of martingale differences with uniformly bounded variances—Bollerslev and Wooldridge [6] establish a similar result when the Gaussian likelihood is employed.

Remark. The LAN condition is straightforward to verify when Cramer-like differentiability conditions are assumed on the log-likelihood function—see Lind and Roussas [26]. We establish this result under weaker smoothness conditions.

We now outline how Theorem 1 is proved. The i.i.d. location scale model considered in Section 2.1 satisfies the LAN condition. In this case, joint quadratic mean differentiability of the square root of the likelihood ratio is sufficient for the LAN condition to hold, as is discussed in Appendix B. To show that this condition holds for the ARCH model requires a conditioning argument.

Let \(h = (h_{\beta}^T, h_{\alpha}^T, h_{\gamma}^T)^T\), then

\[
\varepsilon_T(\theta_T) = \frac{(\varepsilon_T + \delta_{T, t})}{(1 + \eta_{T, t})^{1/2}},
\]

where

\[
\delta_{T, t} = (\beta_T - \beta_0)^T x_t / \sigma_t(\theta_0); \quad \eta_{T, t} = (\sigma_T^2(\theta_T) - \sigma_T^2(\theta_0)) / \sigma_T^2(\theta_0).
\]
Substituting for \( \sigma^2_T(\theta) \) we obtain

\[
\delta_{T,t} = T^{-1/2} \sigma_t^{-1} h_t^T x_t,
\]

\[
\eta_{T,t} = \sigma_t^{-2} \left\{ a_T - a_0 + \sum_{j=1}^P \left[ (c_{jT} - c_{j0}) e_{T-j} \sigma_{T-j}^2 - 2c_{j0} x_{T-j}^T (\beta_T - \beta_0) e_{T-j} \sigma_{T-j} \\
+ c_{j0} (\beta_T - \beta)^T x_{T-j} x_{T-j}^T (\beta_T - \beta_0) \\
- 2(c_{jT} - c_{j0}) x_{T-j}^T (\beta_T - \beta_0) e_{T-j} \sigma_{T-j} \\
+ (c_{jT} - c_{j0}) (\beta_T - \beta_0)^T x_{T-j} x_{T-j}^T (\beta_T - \beta_0) \right] \right\},
\]

where \( \sigma_t^2 = \sigma_t^2(\theta_0) \), \( \epsilon_t = \epsilon_t(\theta_0) \).

Both \( \delta_{T,t} \) and \( \eta_{T,t} \) depend only on the regressors and on the past. In addition, we show in Lemma 1.2 of Appendix B that

\[
\sum_{t=1}^T (\eta_{T,t}^2 + \delta_{T,t}^2) < c < \infty; \quad \max_{1 \leq t \leq T} (\eta_{T,t}^2 + \delta_{T,t}^2) \leq k(T) = 0,
\]

where \( k(T) \) is a deterministic sequence. Thus \( \epsilon_t(\theta_T) \) is close to \( \epsilon_t \). Therefore, the log-likelihood ratio should be well behaved in a neighborhood of the true parameter value.

Let \( \{ \mathcal{T}_s \} : 1 \leq s \leq \infty \) be the increasing family of sigma fields such that \( \mathcal{T}_s = \{ x_t, d_{t-1}, d_{t-2}, \ldots, d_0 \} \), where \( d_t = (y_t, x_t^T)^T \). For convenience sake, we define the following quantities

\[
X_{T,t} = \left[ f(\epsilon_t(\theta_T)) \sigma_t(\theta_0) \over f(\epsilon_t(\theta_0)) \sigma_t(\theta_T) \right]^{1/2} - 1; \quad Z_{T,t} = -\frac{1}{2} T^{-1/2} h_T^T \Delta_t \psi(\epsilon_t),
\]

where \( \Delta_t = (\bar{\sigma}_t^{-1} x_t^T, 0, \ldots, 0)^T \), \( \Delta_2 = (\bar{W}_T^T 1, \bar{\epsilon}_T^T)^T \), \( \Delta_t = (\Delta_{1t}, \Delta_{2t}) \), and \( \psi = (\psi_1, \psi_2)^T \). The i.i.d. vector \( \psi(\epsilon) \) has diagonal covariance matrix \( I \), where \( I = \text{diag}(I_1(f), I_2(f)) \), while the uniformly bounded \( K + P + 1 \) by 2 random matrix \( \Delta_t \) depends only on the past and on the nonstochastic \( x_t \)'s, and therefore is independent of \( \psi(\epsilon) \). The random variable \( Z_{T,t} \) is the (total) quadratic mean derivative of \( X_{T,t} \).

The following proposition is given in Swensen [43]; we verify these conditions in Appendix B.

**PROPOSITION 1.** Assume that the following conditions are satisfied. Then the conclusions of Theorem 1 hold.

1. \( \Sigma_{t=1}^T E(X_{T,t} - Z_{T,t}) = 0; \)
2. \( \sup_T E|\Sigma_{t=1}^T Z_{T,t}| < \infty; \)
3. \( \max_{1 \leq t \leq T} E|Z_{T,t}| < \infty; \)
4. \( \Sigma_{t=1}^T Z_{T,t} \toP h^T J_{88}(\theta_0) h > 0; \)
5. \( \Sigma_{t=1}^T E[Z_{T,t}^2 1(|Z_{T,t}| > \frac{1}{2}) |\mathcal{T}_{t-1}] \toP 0; \)
6. \( E[Z_{T,t} |\mathcal{T}_{t-1}] = 0. \)
Therefore, the LAN property holds for the ARCH model (1.1) and (1.3). This property has two consequences. Fabian and Hannan [14] show that if the log-likelihood ratio satisfies the LAN condition, the Local Asymptotic Minimax bound is achieved by estimators equivalent to the MLE. We discuss this further in the next section.

A second consequence of Theorem 1 is that the sequence of probability measures $P_{T,\theta_T}$ and $P_{T,\theta_0}$ are contiguous in the sense of Roussas [35], Definition 2.1, p. 7. This means that we can interchange the two measures when we make statements about convergence to zero in probability: For any event $A$, we have $P_{T,\theta_T}(A) = 0$ if and only if $P_{T,\theta_0}(A) = 0$. The estimators we consider are constructed from OLS residuals. The significance of the contiguity property is that it enables us to proceed, in many respects, as if we had the true errors instead of these residuals.

**5. ESTIMATION OF $\theta$ WHEN THE ERROR DENSITY IS KNOWN**

Subject to regularity conditions, the MLE of $\theta_0$ when $f$ is known is $\sqrt{T}$ consistent asymptotically normal with covariance matrix $J_{\theta\theta}(\theta_0)^{-1}$. In this section we verify that a two-step estimator based on an initial $\sqrt{T}$ consistent estimator is asymptotically equivalent to the MLE, and is therefore efficient. The precise notion of efficiency that is appropriate here is the Locally Asymptotically Minimax (LAM) criterion of Fabian and Hannan [14] to which paper we refer the reader for a proper definition of this concept. This property is not violated by “superefficient” estimators, unlike the Cramer-Rao lower bound, see Hajek [16]. An alternative efficiency property is that the MLE has the minimal covariance matrix among all uniformly asymptotically normal estimators.

We make the following definition:

**DEFINITION.** A sequence of estimates, $\tilde{\theta}_T$, of $\theta_0$ is asymptotically efficient if it is asymptotically equivalent to the MLE, that is,

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = J_{\theta\theta}(\theta_0,f)^{-1}S_{T\theta}(\theta_0) + o_p(1).$$

For technical reasons we shall restrict ourselves to discretized estimators:

**DEFINITION.** For any sequence of estimators $\tilde{\theta}_T$, define the discretized estimator $\bar{\theta}_T$ to be the nearest vertex of \{ $\theta: \theta = n^{-1/2}(i_1, i_2, \ldots, i_p), i_j$ integers $\}. $

This restriction was employed by Le Cam [25], Bickel [2], and Kreiss [21]. The reason for introducing this concept is that using discretized estimators, we can establish the validity of the Newton-Raphson type estimators without introducing additional differentiability or boundedness assumptions. Kreiss [21], Lemma 4.4, establishes that for any sequence of random variables $q_T(\theta)$, if $q_T(\theta_T) = o_p(1)$, where $|\sqrt{T}(\theta_T - \theta_0)| \leq c$ for some constant
c > 0, then \( q_T(\tilde{\theta}_T) = o_p(1) \) for any discrete and \( \sqrt{T} \) consistent estimator \( \tilde{\theta}_T \). Therefore, we can restrict our attention to nonstochastic sequences \( \theta_T \).

We now consider estimation of \( J_{\theta \theta} \). There are a number of possible consistent estimators: for example, the outer product of the sample scores \( J_{T\theta}(\tilde{\theta}_T) \), where \( \tilde{\theta}_T \) is any discrete and \( \sqrt{T} \) consistent estimator of \( \theta \). Alternatively, we can exploit the known structure of \( J_{\theta \theta} \). Let \( J_{T\theta}(\tilde{\theta}_T) \) be given by (4.1) with \( M(0) \) replaced by \( M_T(0) \), \( \gamma \) replaced by \( \tilde{\gamma}_T \), and \( \tilde{g}_j \) replaced by \( \hat{\tilde{g}}_j \), where

\[
\hat{\tilde{g}}_j = T^{-1} \sum_{t=1}^{T} g_{ij}(\tilde{\theta}_T),
\]

for \( \tilde{\theta}_T \) a \( \sqrt{T} \) consistent discrete estimator of \( \theta_0 \).

To establish the efficiency of our Newton-Raphson estimator defined below we need to establish that our estimator of \( J_{\theta \theta} \) is consistent. This is the content of the following theorem which is proved in Appendix B:

**THEOREM 2.** Let \( \tilde{\theta}_T \) be a discrete \( \sqrt{T} \) consistent estimator of \( \theta \). Then \( J_{T\theta}(\tilde{\theta}_T, f) \) is consistent. 

We also establish that the following asymptotic linearity holds so that we can approximate the estimator by a function linear in i.i.d. random variables.

**THEOREM 3 (ASYMPTOTIC LINEARITY).** Assume that Assumptions 1-7 hold and let \( \theta_T = \theta_0 + T^{-1/2} h \), for any \( h \in \mathbb{R}^{K+P+1} \). Then

\[
S_{T\theta}(\theta_T) - S_{T\theta}(\theta_0) = -J_{\theta\theta}(\theta_0, f) \sqrt{T}(\theta_T - \theta_0) + o_p(1).
\]

This is proved in Appendix B. We are now able to establish the main result of this section.

**THEOREM 4.** Let \( \tilde{\theta}_T \) be a discrete and \( \sqrt{T} \) consistent estimator of \( \theta_0 \), and assume that Assumptions 1-7 hold. Let

\[
\tilde{\theta}_T = \theta_T + T^{-1/2} J_{T\theta}(\tilde{\theta}_T, f)^{-1} S_{T\theta}(\tilde{\theta}_T).
\]

Then \( \tilde{\theta}_T \) is efficient.

Therefore, the linearized MLE of \( \theta \) is asymptotically efficient for a very broad class of densities \( f \). Theorem 4 follows because

\[
\sqrt{T}(\tilde{\theta}_T - \theta_0) = \sqrt{T}(\theta_T - \theta_0) + J_{T\theta}(\tilde{\theta}_T, f)^{-1} S_{T\theta}(\tilde{\theta}_T) = \sqrt{T}(\theta_T - \theta_0) + J_{\theta\theta}(\theta_0)^{-1} S_{T\theta}(\tilde{\theta}_T) + o_p(1)
\]

by Theorem 2

\[
= \sqrt{T}(\theta_T - \theta_0) + J_{\theta\theta}(\theta_0)^{-1}[S_{T\theta}(\theta_0) - J_{\theta\theta}(\theta_0)\sqrt{T}(\tilde{\theta}_T - \theta_0)] + o_p(1)
\]

by Theorem 3

\[
= J_{\theta\theta}(\theta_0)^{-1} S_{T\theta}(\theta_0) + o_p(1).
\]
The results of Theorem 4 complement the existing asymptotic theory for parametric GARCH models described in Weiss [44], Bollerslev and Woolridge [6], and Lumsdaine [27]. These authors establish asymptotic theory for estimators derived from Gaussian PML and least-squares criteria.

6. ESTIMATION OF $\theta$ WHEN THE ERROR DENSITY IS UNKNOWN

We have assumed up to now that the error density is known. We now relax this assumption and construct an estimator that utilizes a consistent estimator of the unknown density.

The first problem we must face is that $\alpha$ and $f$ cannot be separately identified. We can either fix $\alpha$ and let $f$ be unrestricted, or we can estimate $\alpha$ and rescale our estimate of $f$ so that it has unit variance. We assume that $\alpha$ is 0, and is therefore not estimated. We redefine $\theta$ so that $\theta = (\beta^T, \gamma^T)^T \in \mathbb{R}^{P+K}$.

For convenience, we estimate the unknown score function using the kernel method with a normal density function. Undoubtedly, other kernels could be used, and indeed other nonparametric estimation techniques, such as nearest neighbor, splines, or penalized likelihood. See Bickel and Ritov [3].

For any $b = b(T)$, let $\phi(x; b)$ denote the density function of a $N(0, b(T))$ random variable evaluated at $x$. For any $\theta$, we estimate the symmetric error density $f$ by the leave-one-out estimate

$$\hat{f}_{b,t}(x; \theta) = \frac{1}{2(T-1)} \sum_{s=1}^{T} \left[ \phi(x + \varepsilon_s(\theta); b) + \phi(x - \varepsilon_s(\theta); b) \right]$$

$$t = 1, 2, \ldots, T.$$  

This estimator of $f$ is symmetric by construction. As in Bickel [2] and Kreiss [22], we trim out excessive contributions to our estimator. We estimate $\psi_1$ by $\hat{\psi}_{T,t}$, where

$$\hat{\psi}_{T,t}(x; \theta) = \hat{f}_{b,t}(x; \theta) \quad \text{if} \quad \begin{cases} \hat{f}_{b(T),t}(x; \theta) \geq d_T \\
|\hat{f}_{b(T),t}(x; \theta)| \leq c_T \hat{f}_{b(T),t}(x; \theta) \\
|x| \leq e_T \end{cases}$$

$$= 0 \quad \text{else.}$$

We also define $\hat{I}_{T1}(\theta, \hat{f})$, where

$$\hat{I}_{T1}(\theta, \hat{f}) = T^{-1} \sum_{t=1}^{T} \hat{\psi}_{T,t}(\epsilon_t(\theta); \theta)^2; \hat{I}_{T2}(\theta, \hat{f})$$

$$= T^{-1} \sum_{t=1}^{T} [\epsilon_t(\theta) \hat{\psi}_{T,t}(\epsilon_t(\theta); \theta) + 1]^2.$$
The sample scores are estimated by $\hat{S}_{T\beta}(\theta) = \hat{S}_{T\beta 1}(\theta) + \hat{S}_{T\beta 2}(\theta)$, and $\hat{S}_{T\gamma}(\theta)$, where

$$\hat{S}_{T\beta 1}(\theta) = -T^{-1/2} \sum_{t=1}^{T} \sigma_t(\theta)^{-1} x_t \dot{\psi}_{T,t}(\epsilon_t; \theta);$$

$$\hat{S}_{T\beta 2}(\theta) = -T^{-1/2} \sum_{t=1}^{T} [\epsilon_t(\theta) \dot{\psi}_{T,t}(\epsilon_t; \theta) + 1] \tilde{W}_t(\theta);$$

$$\hat{S}_{T\gamma}(\theta) = -\frac{1}{2} T^{-1/2} \sum_{t=1}^{T} [\epsilon_t(\theta) \dot{\psi}_{T,t}(\epsilon_t; \theta) + 1] [\tilde{v}_t(\theta) - \hat{v}(\theta)],$$

while $\hat{v}(\theta)$ is estimated by $\hat{v}(\theta) = T^{-1} \sum_{t=1}^{T} \tilde{v}_t(\theta)$. In proving Theorem 5 and 6 below, we also utilize a form of sample splitting for $\hat{S}_{T\gamma}(\theta)$ similar to that contained in Bickel [2]. This is purely for technical convenience and is not recommended for applications.

We require the bandwidth and trimming sequences to satisfy the following assumption:

Assumption 8. Assume that $b(T)$, $c(T)$, $d(T)$, and $e(T)$ satisfy

1. $b(T), d(T) = 0$, $c(T), e(T) = \infty$,
2. $b(T)c(T) = 0$, $Tb(T)^3c(T)^{-2}e(T)^{-2} = \infty$.

The additional restriction on the bandwidth sequence is required when estimating the score function $\dot{\psi}_2(\cdot)$. With these conditions, we establish the following theorem in Appendix B.

**THEOREM 5.** Let $\theta_T = \theta_0 + T^{-1/2} h$, for any $h \in \mathbb{R}^{K+P}$, and assume that Assumptions 1–8 hold. Then

$$\hat{S}_{T\theta}(\theta_T) - S_{T\theta}(\theta_T) = o_P(1).$$

Furthermore, $\hat{I}_{Tj}(\hat{\theta}_T, f)$ are consistent estimators of $I_j(f)$ for $j = 1, 2$. ■

The information bound for $\gamma$ in the presence of unknown $\alpha$ is $J_{\gamma\gamma}^{*-1}$, where

$$J_{\gamma\gamma}^{*} = J_{\gamma\gamma} - J_{\alpha\gamma} J_{\alpha\alpha}^{-1} J_{\alpha\gamma} = \frac{1}{4} (\hat{g}_{\gamma\gamma} - \hat{g}_{\alpha\gamma} \hat{g}_{\alpha\gamma}^{T}) I_2(f) = \frac{1}{4} \hat{g}_{\gamma\gamma}^{*} I_2(f).$$

We therefore estimate $J_{\beta\beta}$ and $J_{\gamma\gamma}^{*}$ by

$$\hat{J}_{T\beta\beta}(\hat{\theta}_T, \hat{f}) = \{ \hat{g}_{\beta\beta 1} \hat{I}_{T1}(\hat{\theta}_T, \hat{f}) + \hat{g}_{\beta\beta 2} \hat{I}_{T2}(\hat{\theta}_T, \hat{f}) \} M_T(0),$$

$$\hat{J}_{T\gamma\gamma}(\hat{\theta}_T, \hat{f}) = \frac{1}{4} \hat{I}_{T2}(\hat{\theta}_T, \hat{f}) \hat{g}_{\gamma\gamma}^{*}.$$

We now establish the main result of the paper:

**THEOREM 6.** Assume that Assumptions 1–7 hold. Furthermore, let $\hat{\theta}_T$ be a discretized $\sqrt{T}$ consistent estimator of $\theta$. Let

$$\hat{\theta}_T = \bar{\theta}_T + T^{-1/2} \hat{J}_{T\theta}(\bar{\theta}_T, \hat{f})^{-1} \hat{S}_{T\theta}(\bar{\theta}_T),$$
Then

1. \( \hat{J}_{\theta_0}(\hat{\theta}_T, \hat{f}) = J_{\theta_0}(\theta_0, f) + O_P(1) \),
2. \( \sqrt{T}(\theta_T - \theta_0) = J_{\theta_0}(\theta_0, f)^{-1}S_T(\theta_0) + O_P(1) \).  

Therefore,

\[ \sqrt{T}(\hat{\theta}_T - \theta_0) \sim N(0, J_{\theta_0}(\theta_0, f)^{-1}) \]

for all densities \( f \) that satisfy our conditions. Furthermore, \( \hat{J}_{\theta_0}(\hat{\theta}_T, \hat{f})^{-1} \) is a consistent estimator of the asymptotic variance of the adaptive estimator which can be used to form confidence intervals or carry out hypothesis tests.

7. CONCLUSIONS

We have shown how to construct estimates of the identifiable parameters in an ARCH model when the error density is of unknown shape. We have shown that our estimates are adaptive; they have the same asymptotic distribution as the MLE based on the true density. The only substantive restrictions we require on the error density is that it be symmetric about zero.

We expect that a number of extensions of these results are possible. First, the assumption of symmetry could be relaxed as in Bickel [2] and Kreiss [22], although it is not known whether \( \gamma \) is adaptively estimable in this case. Second, an extension to the GARCH model of Bollerslev [5] should be straightforward following the results of Lumsdaine [27]. It should also be possible to allow the bandwidth parameter \( b(T) \) to be data dependent justifying standard cross validation methods for bandwidth choice. Finally, the conditions on our regressors could no doubt be relaxed to allow for trending regressors as well as to include lagged dependent variables.

Bickel's orthogonality condition also holds for the exponential ARCH model defined in (2.1), provided \( r(\cdot) \) is a symmetric function. In this case, we should be able to obtain adaptive estimates of the identifiable parameters, although it remains to provide initial estimates of \( \gamma \) that are \( \sqrt{T} \) consistent for any error density. Furthermore, establishing stationarity of the process for given \( r(\cdot) \) is not a trivial problem.

We employed a number of techniques to establish the asymptotic theory of our estimator: sample splitting, trimming, and discretization. In practice, it may be necessary to trim out the score function estimates, but it is generally agreed—see Hsieh and Manski [20] and Bickel [2]—that sample splitting is unnecessary and undesirable as far as implementing the procedure is concerned.

Our results appear to contradict the simulation evidence of Engle and Gonzalez-Rivera [13] who found a substantial information loss when going from the MLE to the semiparametric estimator. In their parameterization, the ARCH/GARCH variance parameters all contain information about scale. Since knowledge of the error density conveys valuable information about overall scale, one does indeed suffer an information loss when estimating...
these parameters—see Steigerwald [41]. However, the relative effects that are captured by our parameterization are adaptively estimable.

NOTES

1. Another example of this phenomenon is the linear regression model with intercept when the errors are allowed to be asymmetric about zero. At first blush, the scores for the slope parameters are not orthogonal to the tangent space for $f$. Bickel [2] shows that one must first project the slope scores orthogonally to the scores for the intercept. The identifiable parameters—the slopes—are adaptively estimable in this case.

2. Lumsdaine [27] also establishes the positive definiteness of $J_{00}(\theta_0)$. She dispenses with Assumption 5(3) at the cost of strengthening Assumption 4(2).

REFERENCES

27. Lumsdaine, R.L. Asymptotic properties of the quasi-maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models. Manuscript. Harvard University, 1990.
37. Sampson, M. A stationarity condition for the GARCH(1,1) process. Department of Economics, Concordia University (1988).
APPENDIX A

Consider the exponential ARCH model

$$\log [\sigma_t^2] = \alpha + \sum_{j=1}^{p} \gamma_j r(\epsilon_{t-j} \sigma_{t-j}),$$

where $r(\cdot)$ is a known function. This specification differs from that in Nelson [30] in that the innovations are

$$\epsilon_{t-j} \sigma_{t-j} = y_{t-j} - \beta^T x_{t-j},$$

which do not depend on $\alpha$ or $\gamma$. The scores in this model are

$$\dot{t}_{T\beta} = -\sum_{t=1}^{T} [\sigma_t(\theta)^{-1} x_t \dot{\psi}_1(\epsilon_t(\theta)) + \bar{W}_t(\theta) \psi_2(\epsilon_t(\theta))] = \dot{t}_{T\beta 1} + \dot{t}_{T\beta 2},$$

where

$$\bar{W}_t(\theta) = -\frac{1}{2} \sum_{j=1}^{p} \gamma_j x_{t-j} r'(y_{t-j} - \beta^T x_{t-j}),$$

and

$$\dot{t}_{T\gamma} = -\frac{1}{2} \sum_{t=1}^{T} \psi_2(\epsilon_t(\theta)) r_t, \quad \dot{t}_{T\alpha} = -\frac{1}{2} \sum_{t=1}^{T} \psi_2(\epsilon_t(\theta)), $$

where $r_t = (r_1, r_2, \ldots, r_P)^T$, and

$$r_j(\theta) = r(y_{t-j} - \beta^T x_{t-j}), \quad j = 1, 2, \ldots, P.$$

Provided $E[r'(\epsilon, \sigma)] = 0$, the scores for $\beta$ are orthogonal to those for $\alpha$ and $\gamma$, and to any symmetric function of $\epsilon$. In this case, the efficient score function for $\gamma$ in the presence of unknown $\alpha$ is

$$\dot{t}_{T\gamma}^* = -\frac{1}{2} \sum_{t=1}^{T} \psi_2(\epsilon_t(\theta))(r_t - E[r_t]),$$

which is orthogonal to any symmetric function of $\epsilon$. In this case, $\alpha$ is unidentifiable when $f$ is unknown, while both $\beta$ and $\gamma$ are adaptively estimable provided $\sqrt{T}$ consistent initial estimates of $\gamma$ can be found.

APPENDIX B

Proof of Theorem 1. Our treatment is very similar to Swensen [42,43] and Steigerwald [40]. The main difference arises in verifying the quadratic mean differentiability of the log-likelihood ratio. We discuss this point below.
We now verify Conditions (1)–(6) given in Proposition 1. Recall that

\[ Z_{T,t} = T^{-1/2} \{ \pi_{1t} \psi_1(\varepsilon_t) + \pi_{2t} \psi_2(\varepsilon_t) \}, \]

where \( \pi_{jt} \) depend linearly on the fixed regressors \( \{x_{i-j}, j = 1, 2, \ldots, P\} \) and on the bounded random variables \( \sigma_{t-1}, \bar{v}_t, \) and \( \{ \bar{w}_{t-j}, J = 1, 2, \ldots, P\} \), which are all measurable with respect to \( \mathcal{F}_{t-1} \).

Condition (6) is satisfied by Assumption 1. Condition (2) holds by virtue of Assumptions 7, 2, and 4. For example,

\[
T^{-1} \sum_{t=1}^{T} h_{ij} x_{it}^T h_{ij} E[\sigma_{t-2} \psi_1(\varepsilon_t)^2] \leq E[\sigma_{t-2}^2] I_1(f) T^{-1} \sum_{t=1}^{T} h_{ij} x_{it}^T h_{ij},
\]

which is bounded by Assumption 7.

Conditions (3), (4), and (5) can be verified exactly as in Swensen [43]. Without loss of generality, \( \pi_{jt} \) can be treated as deterministic constants obeying Lemma 1.2 below, since the bounded random variables on which they depend can be factored out. Then Condition (3) follows, since in particular

\[
T^{-1/2} \max_{1 \leq t \leq T} |\pi_{jt} \psi_2(\varepsilon_t)| \overset{p}{\to} 0,
\]

because

\[
\text{pr} \left[ T^{-1/2} \max_{1 \leq t \leq T} |\pi_{jt} \psi_2(\varepsilon_t)| > \delta \right] \leq \delta^{-2} T^{-1} \sum_{t=1}^{T} \pi_{jt}^2 E[\psi_2(\varepsilon_t)^2] 1( |\pi_{jt} \psi_2(\varepsilon_t)| > \delta \sqrt{T} )
\]

by Dvoretzky’s inequality, see Hall and Heyde [19], Lemma 2.5. This latter quantity tends to zero with T. Condition (4) follows by (a) and (b) below, where

(a) \( T^{-1} \sum_{t=1}^{T} \pi_{jt}^2 \psi_j(\varepsilon_t)^2 \overset{p}{\to} V_j > 0. \)

This follows because \( T^{-1/2} \sum_{t=1}^{T} \pi_{jt} \psi_j(\varepsilon_t) \) is asymptotically normal—see Swensen [43]. Asymptotic normality is itself a consequence of the following negligibility condition

\[
\left[ \sum_{t=1}^{T} \pi_{jt}^2 \right]^{-1} \max_{1 \leq t \leq T} \pi_{jt}^2 = 0,
\]

which is satisfied by Lemma 1.2 below. The constants \( V_j \) are readily calculated.

(b) \( T^{-1} \sum_{t=1}^{T} \pi_{1t} \psi_1(\varepsilon_t) \psi_2(\varepsilon_t) \overset{p}{\to} 0. \)

This follows since

\[
\text{pr} \left[ T^{-1} \sum_{t=1}^{T} \pi_{1t} \pi_{2t} \psi_1(\varepsilon_t) \psi_2(\varepsilon_t) > \delta \right] \leq T^{-2} \sum_{t=1}^{T} \pi_{1t}^2 \pi_{2t}^2 E[\psi_1(\varepsilon_t)^2 \psi_2(\varepsilon_t)^2],
\]

by Markov’s inequality.

\[
E[\psi_1(\varepsilon_t)^2 \psi_2(\varepsilon_t)^2] \leq E[\psi_1(\varepsilon_t)^4]^{1/2} E[\psi_2(\varepsilon_t)^4]^{1/2} < \infty,
\]

by Cauchy-Schwarz and Assumption 4(3).
Condition (5) can be verified exactly as in Swensen [43].

We now verify Condition (1). First, we need some background on quadratic mean differentiability. Let $\varepsilon$ be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$, with Lebesgue density $f$. Then define the stochastic process $\xi(\varepsilon; \tau) = (\tilde{\xi}_1, \tilde{\xi}_2)^T$ on $(\Omega, \mathcal{F}, P)$, where $\tau = (\delta, \eta)$, and

$$
\xi(\varepsilon; \tau) = \left[ \frac{f((\varepsilon + \delta)/(1 + \eta)^{1/2})}{f(\varepsilon)(1 + \eta)^{1/2}} \right]^{1/2}.
$$

We verify the following lemma below:

**LEMMA 1.1.** Assume that Assumptions 1 and 2 hold. Then $\xi(\varepsilon; \tau)$ is jointly quadratic mean differentiable at any $(\delta, \eta)$ where $\eta > -1$. In other words, there exists a vector process $d\xi(\varepsilon; \tau) = (d\xi_1, d\xi_2)^T$ such that for any $u = (m, s)^T = 0$,

$$
\lim_{m, s \to 0} E[(m^2 + s^2)^{-1/2}[\xi(\varepsilon; \delta + m, \eta + s) - \xi(\varepsilon; \delta, \eta) - u^T d\xi(\varepsilon, \delta, \eta)]^2 = 0,
$$

independently of the path $(m, s) = 0$.

This concept was introduced in Le Cam [25]. Joint quadratic mean differentiability implies marginal q.m.d., and we define $d\xi_1(\varepsilon; \tau)$ and $d\xi_2(\varepsilon; \tau)$ as satisfying

$$
\begin{align*}
\lim_{m \to 0} E\{m^{-1}[\xi(\varepsilon; \delta + m, \eta) - \xi(\varepsilon; \delta, \eta)] - d\xi_1(\varepsilon; \delta, \eta)]^2 &= 0, \\
\lim_{s \to 0} E\{s^{-1}[\xi(\varepsilon; \delta, \eta + s) - \xi(\varepsilon; \delta, \eta)] - d\xi_2(\varepsilon; \delta, \eta)]^2 &= 0,
\end{align*}
$$

where

$$
\begin{align*}
d\xi_1(\varepsilon; \tau) &= \xi(\varepsilon; \tau) \frac{f'}{f} \left( \frac{\varepsilon + \delta}{(1 + \eta)^{1/2}} \right) (1 + \eta)^{-1/2}; \\
d\xi_2(\varepsilon; \tau) &= -\xi(\varepsilon; \tau) \frac{1}{2} \left\{ \left( \frac{\varepsilon + \delta}{(1 + \eta)^{1/2}} \right) \frac{f'}{f} \left( \frac{\varepsilon + \delta}{(1 + \eta)^{1/2}} \right) + 1 \right\} (1 + \eta)^{-1};
\end{align*}
$$

It remains to show quadratic mean differentiability for the regression model with ARCH errors. We have to show that

$$
\sum_{t=1}^{T} E(X_{T,t} - Z_{T,t})^2 = 0.
$$

We establish (B.1) and (B.2) below which together imply this:

$$
\sum_{t=1}^{T} E(X_{T,t} - Z_{T,t}^*)^2 = 0; \quad \sum_{t=1}^{T} E(Z_{T,t} - Z_{T,t}^*)^2 = 0
$$

where

$$
Z_{T,t}^* = -\frac{1}{2} \xi_{T,t}^T \psi(\varepsilon), \quad \xi_{T,t} = (\delta_{T,t}, \frac{1}{2} \eta_{T,t})^T.
$$

We first show (B.2).
\[ E \sum_{t=1}^{T} (Z_{T,t} - Z_{T,t}^*)^2 \leq qI_2(f) \left( E[\sigma_t^{-4}] \left[ T^{-2} \sum_{t=1}^{T} \sum_{j} \sum_{k} (c_{j0}c_{k0}(h_{\beta}^T x_{T-j})^2(h_{\beta}^T x_{T-k})^2) \right] \right) + 4T^{-1} \sum_{t=1}^{T} \sum_{j} (c_{j0} - c_{jT})^2(x_{T-j}^T h_{\beta})^2 E[\sigma_t^{-4}c_{T-j}^2\sigma_{T-j}^2] + E[\sigma_t^{-4}]T^{-2} \sum_{t=1}^{T} \sum_{j} \sum_{k} (c_{j0} - c_{jT})(c_{k0} - c_{kT})(h_{\beta}^T x_{T-j})^2(h_{\beta}^T x_{T-k})^2 + 2E[\sigma_t^{-4}]T^{-2} \sum_{t=1}^{T} \sum_{j} \sum_{k} (c_{j0} - c_{jT})c_{k0}(h_{\beta}^T x_{T-j})^2(h_{\beta}^T x_{T-k})^2 \right) + O(T^{-1}), \]

where \( q \) is a constant reflecting the number of times we expanded out a sum of squares, while the \( O(T^{-1}) \) remainder term we did not specify arises from the approximations

\[ e^{\alpha_T - \alpha_0} - 1 = \alpha_T - \alpha_0 + O(T^{-1}); \]
\[ c_{jT} - c_{j0} = e^{\alpha_0(\gamma_{jT} - \gamma_{j0})} + \gamma_{j0}(\alpha_T - \alpha_0) + O(T^{-1}). \]

Therefore,

\[ E \sum_{t=1}^{T} (Z_{T,t} - Z_{T,t}^*)^2 = 0, \]

because for example

\[ T^{-2} \sum_{t=1}^{T} \sum_{j} \sum_{k} (c_{j0}c_{k0}(h_{\beta}^T x_{T-j})^2(h_{\beta}^T x_{T-k})^2) = 0 \]

by Lemma 1.2.

We now show (B.1). Since \( \varepsilon_t \) is independent of both \( \delta_{T,t} \) and \( \eta_{T,t} \), we have to show that \( L = \sum_{t=1}^{T} E[L_{T,t}] = 0, \) where

\[ L_{T,t} = \int \left[ \hat{f}(\varepsilon; \delta_{T,t}, \eta_{T,t}) - 1 - \frac{1}{2} \xi_{T,t}^2 \psi(\varepsilon) \right] f(\varepsilon) \, d\varepsilon. \]

We use the following lemma:

**Lemma 1.2.** There is a constant \( c \) and a deterministic sequence \( k(T) = 0 \) such that

\[ \max_{1 \leq t \leq T} (\eta_{T,t}^2 + \delta_{T,t}^2) \leq k(T); \quad \sum_{t=1}^{T} (\eta_{T,t}^2 + \delta_{T,t}^2) < c < \infty. \]

Define the following family of neighborhoods of zero:

\[ B(k) = \{ (\delta, \eta) : (\eta^2 + \delta^2) < k \}. \]
Then,
\[ L \leq \sum_{t=1}^{T} (\delta_{T,t}^2 + \eta_{T,t}^2)^* \sup_{B(k)} \left\{ (\delta^2 + \eta^2)^{-1} \int \left[ \xi(\varepsilon; \delta, \eta) - 1 - \frac{1}{2} \xi^T \psi(\varepsilon) \right]^2 f(\varepsilon) \, d\varepsilon \right\}. \]
Since \( k(T) = 0 \), \( L = 0 \) as required.

**Proof of Theorem 2.** Since \( \{g_{ij}(\theta_0)\}_{t=1}^{T} \) is a bounded stationary ergodic process, we have
\[ T^{-1} \sum_{t=1}^{T} g_{ij}(\theta_0) \xrightarrow{P} g_{ij}(\theta_0) \]
by the ergodic theorem—see, for example, Hall and Heyde [19], p. 281. It is easy to verify that in each case there is a positive constant \( K \) and a neighborhood \( \mathcal{N}_{\theta_0} \) of \( \theta_0 \) on which
\[ T^{-1} \sum_{t=1}^{T} |g_{ij}(\theta_1) - g_{ij}(\theta_2)| \leq K \| \theta_1 - \theta_2 \|, \]
for large \( T \). For example, consider \( g_{ij}(\theta) = \sigma_\tau^{-2}(\theta) \). In this case,
\[ \{\sigma_\tau^{-2}(\theta_T) - \sigma_\tau^{-2}(\theta_0)\} = \sigma_\tau^{-2}(\theta_T) \eta_{T,i}, \]
for any sequence \( \theta_T \) such that \( \sqrt{T}(\theta_T - \theta_0) \) stays bounded. Therefore,
\[ T^{-1} \sum_{t=1}^{T} \{\sigma_\tau^{-2}(\theta_T) - \sigma_\tau^{-2}(\theta_0)\} \leq \| \theta_T - \theta_0 \| T^{-1} \sum_{t=1}^{T} a_t, \]
where \( \{a_t\} \) is a deterministic sequence derived from \( \{\eta_{T,i}\} \), and \( |T^{-1} \sum_{t=1}^{T} a_t| \) is bounded.

It follows that
\[ T^{-1} \sum_{t=1}^{T} (g_{ij}(\tilde{\theta}_T) - g_{ij}(\theta_0)) \xrightarrow{P} 0, \]
where \( \tilde{\theta}_T \) is a discrete \( \sqrt{T} \) consistent estimator of \( \theta_0 \), and \( J_\theta(\theta_T, f) \) is consistent.

**Proof of Theorem 3.** First, we write
\[ S_{T_0}(\theta_T) = S_{T_0}(\theta_0) + J_{\theta_0}(\theta_0, f) \sqrt{T}(\theta_T - \theta_0) = Q_{T_0}(\theta_T) - Q_{T_0}(\theta_0) + Y_{T_0}(\theta_T), \]
where
\[ Q_{T_0}(\theta) = T^{-1/2} \sum_{t=1}^{T} \{s_\theta(\theta) - E_{\theta_0}[s_\theta(\theta) \mid \mathcal{F}_{t-1}]\}, \]
is a standardized sum of martingale differences with respect to \( \mathcal{F}_{t-1} \), and
\[ Y_{T_0}(\theta_T) = T^{-1/2} \sum_{t=1}^{T} \left\{ E_{\theta_0}[s_\theta(\theta_T) \mid \mathcal{F}_{t-1}] + J_{\theta_0}(\theta_0, f) \sqrt{T}(\theta_T - \theta_0) \right\}, \]
where \( E_{\theta_0} [\delta_{\theta_0} (\theta_0) \mid \mathcal{T}_{t-1}] = 0 \). We show

\[
Q_{\mathcal{T}_T}(\theta_T) - Q_{\mathcal{T}_T}(\theta_0) = o_P(1); \tag{B.4}
\]
\[
Y_{\mathcal{T}_T}(\theta_T) = o_P(1), \tag{B.5}
\]
where \( Q_{\mathcal{T}_T} = (Q_{\mathcal{T}_T}, Q_{\mathcal{T}_a}, Q_{\mathcal{T}_y})^T \), and \( Q_{\mathcal{T}_3} = Q_{\mathcal{T}_{31}} + Q_{\mathcal{T}_{32}} \).

**Proof of (B.4).** We establish that \( Q_{\mathcal{T}_3}(\theta_T) - Q_{\mathcal{T}_3}(\theta_0), Q_{\mathcal{T}_a}(\theta_T) - Q_{\mathcal{T}_a}(\theta_0) \), and \( Q_{\mathcal{T}_y}(\theta_T) - Q_{\mathcal{T}_y}(\theta_0) \) are \( o_P(1) \).

By the triangle inequality

\[
E[\|Q_{\mathcal{T}_3}(\theta_T) - Q_{\mathcal{T}_3}(\theta_0)\|^2] 
\leq 2E[\|Q_{\mathcal{T}_{31}}(\theta_T) - Q_{\mathcal{T}_{31}}(\theta_0)\|^2] + 2E[\|Q_{\mathcal{T}_{32}}(\theta_T) - Q_{\mathcal{T}_{32}}(\theta_0)\|^2].
\]

Let

\[
\theta_{T,i} = \sigma_i(\theta_T)^{-1} x_i \psi_1(\varepsilon_i(\theta_T)) - E_{\theta_0} [\sigma_i(\theta_T)^{-1} x_i \psi_1(\varepsilon_i(\theta_T)) \mid \mathcal{T}_{t-1}] - \sigma_i^{-1} x_i \psi_1(\varepsilon_i).
\]

Then

\[
E\left(\|Q_{\mathcal{T}_{31}}(\theta_T) - Q_{\mathcal{T}_{31}}(\theta_0)\|^2\right) \leq T^{-1} \sum_{i=1}^T E[\theta_{T,i} \theta_{T,i}^T],
\]

since \( \forall T \{\theta_{T,i}\} \) is a martingale difference sequence with respect to \( \mathcal{T}_{t-1} \). Furthermore,

\[
T^{-1} \sum_{i=1}^T E[\theta_{T,i} \theta_{T,i}^T] \leq 2\|A\| + 2\|B\|,
\]

where

\[
A = T^{-1} \sum_{i=1}^T x_i x_i^T E_{\theta_0} [\sigma_i(\theta_T)^{-1} \psi_1(\varepsilon_i(\theta_T)) - \sigma_i^{-1} \psi_1(\varepsilon_i)^2],
\]

and

\[
B = T^{-1} \sum_{i=1}^T x_i x_i^T E_{\theta_0} [\sigma_i(\theta_T)^{-1} E[\psi_1(\varepsilon_i(\theta_T)) \mid \mathcal{T}_{t-1}]^2].
\]

We can bound \( \|A\| \) by \( 2\|A_1\| + 2\|A_2\| \), where

\[
A_1 = T^{-1} \sum_{i=1}^T x_i x_i^T E_{\theta_0} [\sigma_i(\theta_T)^{-2} [\psi_1(\varepsilon_i(\theta_T)) - \psi_1(\varepsilon_i)]^2];
\]

\[
A_2 = T^{-1} \sum_{i=1}^T x_i x_i^T E_{\theta_0} [(\sigma_i(\theta_T)^{-1} - \sigma_i^{-1}) \psi_1(\varepsilon_i)]^2.
\]

Since \( \sigma_i(\theta_T)^{-2} \) is eventually bounded from above by some \( \delta < \infty \), we have for large \( T \)

\[
\sigma_i(\theta_T)^{-2} [\psi_1(\varepsilon_i(\theta_T)) - \psi_1(\varepsilon_i)]^2 \leq \delta [\psi_1(\varepsilon_i(\theta_T)) - \psi_1(\varepsilon_i)]^2. \tag{B.6}
\]

Furthermore, \( \psi_1(\varepsilon_i) \) is independent of \( \sigma_i^{-1} \) and \( \sigma_i(\theta_T)^{-1} \), because these latter quantities depend only on the past. Therefore,

\[
E_{\theta_0} [(\sigma_i(\theta_T)^{-1} - \sigma_i^{-1}) \psi_1(\varepsilon_i)]^2 = I_1(f) E_{\theta_0} [(\sigma_i(\theta_T)^{-1} - \sigma_i^{-1})^2]. \tag{B.7}
\]
Together, (B.6) and (B.7) imply that we have to estimate the norms of the following matrices

$$A_1' = T^{-1} \sum_{t=1}^{T} x_t x_t^T E[\psi_1(\varepsilon(\theta_T)) - \psi_1(\varepsilon_t)]^2;$$

$$A_2' = I(f)T^{-1} \sum_{t=1}^{T} x_t x_t^T E[\sigma_\varepsilon(\theta_T)^{-1} - \sigma_{\varepsilon_t}^{-1}]^2.$$

Since

$$\max_{1 \leq t \leq T} E_{\theta_0} [\sigma_\varepsilon(\theta_T)^{-1} - \sigma_{\varepsilon_t}^{-1}]^2 = 0, \quad \text{(B.8)}$$

by Lemma 1.2, and

$$\mathbb{E} \left\{ \int \psi_1((\varepsilon + \delta_{T,t})/(1 + \eta_{T,t})^{1/2}) - \psi_1(\varepsilon) \right\}^2 f(\varepsilon) \, d\varepsilon \right\} \leq \sup_{B(k)} \left\{ \int \psi_1((\varepsilon + \delta)/(1 + \eta)^{1/2}) - \psi_1(\varepsilon) \right\}^2 f(\varepsilon) \, d\varepsilon = 0, \quad \text{(B.9)}$$

then $\|A_1'\|, \|A_2'\| = 0$.

Similarly, $B = 0$, because

$$T^{-1} \sum_{T=1}^{T} E_{\theta_0} \left[ \mathbb{E}[\psi_1(\varepsilon(\theta_T)) | T_{t-1}] \right]^2 \leq \sup_{B(k)} \left\{ \int \psi_1((\varepsilon + \delta)/(1 + \eta)^{1/2})f(\varepsilon) \, d\varepsilon \right\}^2 = 0$$

by Assumption 3(4).

Since $\tilde{\psi}_T(\theta_T)$, $\tilde{\psi}_1(\theta_T)$, and $\sigma_\varepsilon(\theta_T)^{-1}$ are eventually bounded, the same reasoning can be applied to show that

$$E[\|Q_{T^2,1}(\theta_T) - Q_{T^2,2}(\theta_0)\|^2] = 0,$$

$$E[\|Q_{T^2}(\theta_T) - Q_{T^2}(\theta_0)\|^2] = 0,$$

$$E[\|Q_{T^2}(\theta_T) - Q_{T^2}(\theta_0)\|^2] = 0.$$

**Proof of (B.5).** We first examine the terms due to $\beta$. We have to show that

$$T^{-1/2} \sum_{t=1}^{T} E_{\theta_0} [\delta_{tj}(\theta_T) | \Psi_{t-1}] + J_{\beta j} \sqrt{T}(\beta_T - \beta) = o_p(1), \quad j = 1, 2.$$  

This amounts to showing that

$$T^{-1/2} \sum_{t=1}^{T} E[\delta_{tj}(\theta_T)] = J_{\beta j} h_{\beta} + o(1), \quad j = 1, 2.$$  

We examine $Y_{T,1}$, the argument for $Y_{T,2}$ is the same, and is omitted. We have

$$E_{\theta_0} [\delta_{tj}(\theta_T) | \Psi_{t-1}] = \sigma_\varepsilon(\theta_T)^{-1} x_t \int \psi_1((\varepsilon + \delta_{T,t})/(1 + \eta_{T,t})^{1/2}) f(\varepsilon) \, d\varepsilon.$$
Therefore,
\[
\left| T^{-1/2} \sum_{t=1}^{T} E \left[ (\sigma_t(\theta_T)^{-1} x_t \int \phi_t((\varepsilon + \delta_{T, t})/(1 + \eta_{T, t})^{1/2}) f(\varepsilon) \, d\varepsilon \right. \right.
\]
\[
\left. - T^{-1/2} \sigma_t^{-2} x_t x_t^T h_{\beta} I_1(f) \right] \right|
\]
\[
\leq \left| T^{-1/2} \sum_{t=1}^{T} E \left[ (\sigma_t(\theta_T)^{-1} x_t \int \phi_t((\varepsilon + \delta_{T, t})/(1 + \eta_{T, t})^{1/2}) f(\varepsilon) \, d\varepsilon \right. \right.
\]
\[
\left. - T^{-1/2} (\sigma_t(\theta_T)^{-2} x_t x_t^T h_{\beta} I_1(f)) \right]
\]
\[
+ \left| T^{-1} \sum_{t=1}^{T} (\sigma_t^{-2}(\theta_T) - \sigma_t^{-2}) x_t x_t^T h_{\beta} I_1(f) \right].
\]
Using the argument given in Theorem 2,
\[
\left| T^{-1} \sum_{t=1}^{T} (\sigma_t(\theta_T)^{-2} - \sigma_t^{-2}) x_t x_t^T \right| \overset{P}{=} 0.
\]
By the eventual boundedness of \(\sigma_t(\theta_T)^{-1}\), the first term is less than
\[
\left| T^{-1/2} \sum_{t=1}^{T} x_t E \left[ \int \phi_t((\varepsilon + \delta_{T, t})/(1 + \eta_{T, t})^{1/2}) f(\varepsilon) \, d\varepsilon - T^{-1/2} \sigma_t^{-1} x_t x_t^T h_{\beta} I_1(f) \right] \right|
\]
\[
\leq \left| T^{-1/2} \sum_{t=1}^{T} \delta_{T, t} x_t \sup_{B(k)} \delta^{-1} \left[ \int \phi_t((\varepsilon + \delta)/(1 + \eta)^{1/2}) f(\varepsilon) \, d\varepsilon \right.ight.
\]
\[
\left. \left. - T^{-1/2} \sigma_t^{-1} x_t x_t^T h_{\beta} I_1(f) \right] \right|
\]
\[
\leq \left| T^{-1} \sum_{t=1}^{T} x_t x_t^T h_{\beta} \sigma_t^{-1} \sup_{B(k)} \delta^{-1} \left[ \int \phi_t((\varepsilon + \delta)/(1 + \eta)^{1/2}) f(\varepsilon) \, d\varepsilon + I_1(f) \right] \right| \Rightarrow 0
\]
by Assumption 3(2).

**Proof of Lemmas.** Lemma 1.1. Swensen [42] Lemmas 3 and 4, pp. 56–57, establish that Assumptions 1(a) and 2 are sufficient to guarantee that the related process
\[
\tilde{\gamma}(\varepsilon; \tau) = \left[ \frac{f((\varepsilon + \mu)/\sigma)}{\sigma f(\varepsilon)} \right]^{1/2}
\]
is jointly differentiable in quadratic mean \(\psi_\mu, \sigma\) for \(\sigma > 0\). We change variables from \(\sigma\) to \((1 + \eta)^{1/2}\). Since \((1 + x)^{1/2}\) is continuous in \(x\), the result follows.

Lemma 1.2. Assumption 7 implies that
1. \(\max_{1 \leq t \leq T} x_t \left( \sum_{t=1}^{T} x_t x_t^T \right)^{-1} x_t = 0\);
2. \(T^{-1} \max_{1 \leq t \leq T} x_t^T x_t = 0\),
see Wu [46], Lemma 3. Expand out $\delta_{T,t}^2 + \eta_{T,t}^2$ using the triangle inequality. Since $\sigma_t^{-1}$, $\sigma_t^2 \epsilon_{t-j} \sigma_{t-j}$, and $\sigma_t^{-2} \epsilon_{t-j}^2 \sigma_{t-j}^2$ are bounded, they can be factored out of the expression. Then, for example,

$$T^{-2} \max_{1 \leq t \leq T} (h_{h_t}^T x_{t-j})^4 \leq \left[ T^{-1} \max_{1 \leq t \leq T} (h_{h_t}^T x_{t-j})^2 \right]^2 = 0,$$

$$T^{-2} \sum_{t=1}^T (h_{h_t}^T x_{t-j})^4 \leq T^{-1} \max_{1 \leq t \leq T} (h_{h_t}^T x_{t-j})^2 T^{-1} \sum_{t=1}^T (h_{h_t}^T x_{t-j})^2 \Rightarrow 0.$$

The result follows.

**Proof of Theorem 5.** All calculations below are carried out under the measure $P_{T,\theta_0}$, by contiguity the convergence to zero in probability also holds under $P_{T,\theta_0}$.

Now consider

$$E \| \hat{S}_{T3}(\theta_T) - S_{T3}(\theta_T) \|^2 \leq 2E \| \hat{S}_{T31}(\theta_T) - S_{T31}(\theta_T) \|^2 + 2E \| \hat{S}_{T32}(\theta_T) - S_{T32}(\theta_T) \|^2.$$

By construction, $\hat{w}_{T,t}(x; \theta_T)$ is antisymmetric about zero, that is,

$$\hat{w}_{T,t}(-x; \theta_T) = -\hat{w}_{T,t}(x; \theta_T),$$

for all $x$. Therefore, as in Kreiss [21] and Bickel [2], we obtain

$$E \| \hat{S}_{T31}(\theta_T) - S_{T31}(\theta_T) \|^2 = T^{-1} \sum_{t=1}^T x_t x_t^T E \left[ \sigma_t(\theta_T)^{-2} \int \left\{ \hat{w}_{T,t}(x; \theta_T) - \psi_1(x) \right\}^2 f(x) \, dx \right],$$

and because $\sigma_t(\theta_T)^{-2}$ is bounded for large $T$, we can apply directly the results of Kreiss [21]. We have

$$\max_{1 \leq t \leq T} E \int \left\{ \hat{w}_{T,t}(x; \theta_T) - \psi_1(x) \right\}^2 f(x) \, dx = 0.$$

Therefore,

$$E \| \hat{S}_{T31}(\theta_T) - S_{T31}(\theta_T) \|^2 = 0.$$

To establish the same for $E \| \hat{S}_{T32}(\theta_T) - S_{T32}(\theta_T) \|^2$, we must exploit the functional form of $\hat{w}_t$. The estimated scale score $\hat{w}_{T,t}(x; \theta_T) x + 1$ is symmetric in both $x$ and $\epsilon_{t-j}(\theta_T)$ for any $j$, while $\hat{w}_t$ is antisymmetric in $\epsilon_{t-1}(\theta_T)$, that is

$$\hat{w}_t(-\epsilon_{t-1}(\theta_T)) = -\hat{w}_t(\epsilon_{t-1}(\theta_T)).$$

Therefore, the cross products drop out

$$E \left[ \hat{w}_t \hat{w}_T^T \epsilon \epsilon_1 \{ \hat{w}_{T,t}(\epsilon; \theta_T) - \psi_1(\epsilon_1) \} \{ \hat{w}_{T,s}(\epsilon_1; \theta_T) - \psi_1(\epsilon_1) \} \right] = 0,$$

and

$$E \| \hat{S}_{T32}(\theta_T) - S_{T32}(\theta_T) \|^2 = T^{-1} \sum_{t=1}^T E \left[ \hat{w}_t(\theta_T) \hat{w}_t(\theta_T)^T \right] \int \left\{ \hat{w}_{T,t}(x; \theta_T) - \psi_1(x) \right\}^2 x^2 f(x) \, dx.$$

By a minor modification of Kreiss’s [21] arguments, we can establish that
\[
\max_{1 \leq t \leq T} E_{\theta_T} \left[ \int \{ \hat{\psi}_{T,t}(x; \theta_T) - \psi_1(x) \}^2 x^2 f(x) \, dx \right] = 0, \tag{B.10}
\]

via a sequence of standard arguments collected below in Lemmas 5.1–5.5. Therefore, since

\[
T^{-1} \sum_{t=1}^{T} E[\tilde{W}_t(\theta_T)\tilde{W}_t(\theta_T)^T] < M,
\]

for some \( M < \infty \), the result follows.

Notice that Kreiss’s argument does not require any sample splitting. However, when we examine the estimated scores for \( \gamma \), we are unable to exploit symmetry properties and the argument becomes considerably more involved. We adopt a form of sample splitting in order to provide a simple proof. We split the sample into two subsamples.

\[
I_1 = \{t: t = 1, 2, \ldots, T_1\}; \quad I_2 = \{t: t = T_1 + 1, \ldots, T\},
\]

where

\[
T_1(T) = \infty; \quad T_1/T = 0 \quad \text{as} \quad T = \infty.
\]

The first subsample is used to estimate the score function \( \psi_1(x) \), while the remaining observations are used to construct the estimator.

In this case,

\[
\hat{S}_{T_1}(\theta_T) - S_{T_1}(\theta_T) = T^{-1/2} \sum_{t \in I_2} \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\hat{\nu}_t(\theta_T) - \hat{\nu}),
\]

where \( \hat{\nu}(\theta_T) = T^{-1} \sum_{t \in I_2} \hat{\nu}_t(\theta_T) \). For economy of notation, we drop the \( \theta_T \) argument.

We have

\[
\hat{S}_{T_1} - S_{T_1} = T^{-1/2} \sum_{t \in I_2} \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\hat{\nu}_t - \hat{\nu})
\]

\[+ T^{1/2}(\hat{\nu} - \hat{\nu})T^{-1} \sum_{t \in I_2} \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t)),\]

\[= I + II.\]

But \( \hat{\nu}_t - \hat{\nu} \) is zero mean and independent of \( \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t)) \), and hence

\[
E \left[ \left\| T^{-1/2} \sum_{t \in I_2} \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\hat{\nu}_t - \hat{\nu}) \right\|^2 \right]
\]

\[= T^{-1} \sum_{t \in I_2} E[3(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\hat{\nu}_t - \hat{\nu})^2].\]

Since \( \hat{\nu}_t \) is bounded, if Lemma 5.1 holds, then \( I = o_p(1) \) as required. The second term \( II \) is \( o_p(1) \) because

\[
T^{1/2}(\hat{\nu} - \hat{\nu}) = o_p(1), \tag{B.11}
\]

\[
T^{-1} \sum_{t \in I_2} \varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t)) = o_p(1), \tag{B.12}
\]

by (B.10).
We now establish the fundamental property (B.10). Let

\[ f_b(x) = \int \phi(x - y; b)f(y) \, dy; \quad \psi_b(x) = \frac{f_b}{f_b}(x). \]

By repeated addition and subtraction, we get that

\[
\int x^2(\hat{\psi}_{T,t}(x; \theta_T) - \psi_1(x))^2f(x) \, dx
\]

\[
\leq 3 \left\{ \int x^2 \left( \hat{\psi}_{T,t}(x; \theta_T) - \hat{\psi}_{T,t}(x; \theta_T) \left[ \frac{f_b}{f}(x) \right]^{1/2} \right)^2f(x) \, dx
\]

\[
+ \int x^2 \left( \hat{\psi}_{T,t}(x; \theta_T) \left[ \frac{f_b}{f}(x) \right]^{1/2} - \psi_b(x) \left[ \frac{f_b}{f}(x) \right]^{1/2} \right)^2f(x) \, dx
\]

\[
+ \int x^2 \left( \left[ \frac{f_b}{f}(x) \right]^{1/2} \left[ \frac{f_b}{f_b}(x) - \frac{f'}{f^{1/2}}(x) \right] \right)^2f(x) \, dx
\]

\[
= 3 \int x^2 \hat{\psi}_{T,t}(x; \theta_T)^2 \left( f_b(x)^{1/2} - f(x)^{1/2} \right)^2 \, dx
\]

\[
+ 3 \int x^2(\hat{\psi}_{T,t}(x; \theta_T) - \psi_b(x))^2f_b(x) \, dx + 3 \int x^2 \left[ \frac{f_b}{f^{1/2}}(x) - \frac{f'}{f^{1/2}}(x) \right]^2 \, dx.
\]

The second term is essentially a "variance" term, while the remaining terms are "biases."

The following lemmas are, apart from the factor \(x^2\), identical to Lemmas 6.5 and 6.6 proved in Kreiss [21].

**Lemma 5.1.** For each \(x \in \mathbb{R}\), there are constants \(k_0\) and \(k_1\) such that

\[
E_{T,T}[x^2f_b^{-1}(x) \{ \hat{f}_{b,x}(x; \theta_T) - f_b(x) \}^2] \leq \frac{1}{b(T)T} \{ k_0x^2 + k_1x^4T^{-1} \}.
\]

**Lemma 5.2.** For each \(x \in \mathbb{R}\), there are constants \(K_0\) and \(K_1\) such that

\[
E_{T,T}[F_b^{-1}(x)x^2 \{ \hat{f}_{b,x}(x; \theta_T) - f_b(x) \}^2] \leq \frac{1}{b(T)^3T} \{ K_0x^2 + K_1x^4T^{-1} \}.
\]

**Lemma 5.3.** As \(T \to \infty\)

\[
\int x^2 \left[ \frac{f_b}{f^{1/2}}(x) - \frac{f'}{f^{1/2}}(x) \right]^2 \, dx = o(1).
\]

**Proof.** This holds because

\[
\int \psi_b(x)^2f_b(x)x^2 \, dx < \int \psi_1(x)^2f(x)x^2 \, dx < \infty,
\]

since \(I_2 < \infty\). Therefore, we can apply dominated convergence.
LEMMA 5.4. Provided $Tb(T)^3 c(T)^{-2} e(T)^{-2} \Rightarrow 0$,
\[
\max_{1 \leq t \leq T} \mathbb{E}_{\theta_T} \int x^2 [\hat{\psi}_{T,t}(x;\theta_T) - \psi_b(x)]^2 f_b(x) \, dx = o(1).
\]

**Proof.** This is the same as Lemma 6.8 in Kreiss [21] apart from the additional factor of $x^2$ and some constants. We have
\[
[\hat{\psi}_{T,t}(x;\theta_T) - \psi_b(x)]^2 f_b(x) \leq 2 \left[ \frac{\hat{f}_{b,t}(x)}{f_b(x)} - \frac{f_{b,t}(x)}{f_b(x)} \right]^2 f_b(x) + 2 \left[ \frac{\hat{f}_{b,t}(x)}{f_b(x)} - \frac{f_{b,t}(x)}{f_b(x)} \right]^2 f_b(x).
\]
Since
\[
\left[ \frac{\hat{f}_{b,t}(x)}{f_b(x)} - \frac{f_{b,t}(x)}{f_b(x)} \right]^2 f_b(x) = \left[ \frac{\hat{f}_{b,t}(x)}{f_b(x)} \right]^2 \left[ f_{b,t}(x) - f_b(x) \right]^2 f_b(x)^{-1} \leq c(T)^2 \left[ \hat{f}_{b,t}(x) - f_b(x) \right]^2 f_b(x)^{-1},
\]
we can apply Lemmas 5.1 and 5.2 above, after truncating the integral according to the sets\[
A_{T,t} = \{ x; \hat{f}_{b,t}(x;\theta_T) \geq d_T \}; \quad B_{T,t} = \{ x; |x| < e_T \}; \quad C_{T,t} = \{ x; \hat{\psi}_{T,t}(x;\theta_T) \leq c_T \}.
\]
The proof is exactly the same as Kreiss [21].

LEMMA 5.5. Provided $b(T)c(T) = 0$,
\[
\max_{1 \leq t \leq T} \mathbb{E}_{\theta_T} \left\{ \int x^2 [\hat{\psi}_{T,t}(x;\theta_T) - \sqrt{f_b(x)}]^2 \, dx \right\} = o(1).
\]

**Proof.** Since $\hat{\psi}_{T,t}^2 \leq c(T)^2$, this random variable is bounded by
\[
c(T)^2 \int [x^2 [\sqrt{f_b(x)} - \sqrt{f_b(x)}]^2 \, dx] = O(b(T)^2 c(T)^2),
\]
by Bickel [2], Lemma 6.3. It is $o(1)$ provided $b(T)c(T) \Rightarrow 0$. 

\[\boxed{}\]